Frontier-based vs. traditional mutual fund ratings: A first backtesting analysis[☆]

Olivier Brandouy[a], Kristiaan Kerstens[b,*], Ignace Van de Woestyne[c]

[a] *GREThA, UMR CNRS 5113, Université Montesquieu (Bordeaux IV), avenue Lon Duguit, F-33608, Pessac cedex, France*
[b] *CNRS-LEM (UMR 8171), IESEG School of Management, 3 rue de la Digue, F-59000 Lille, France*
[c] *KU Leuven, Research unit MEES, Stormstraat 2, B-1000 Brussels, Belgium*

Abstract

We explore the potential benefits of a series of existing and new non-parametric convex and non-convex frontier-based fund rating models to summarize the information contained in the moments of the mutual fund price series. Limiting ourselves to the traditional mean-variance portfolio setting, we test in a simple backtesting setup whether these efficiency measures fare any better than more traditional financial performance measures in selecting promising investment opportunities. The evidence points to a remarkable superior performance of these frontier models compared to most, but not all traditional financial performance measures.

Keywords: mutual fund rating, DEA, FDH, shortage function, mean-variance portfolio frontier.

1. Introduction

Investors look at mutual fund (MF) performance measures established by rating agencies around the world (e.g., Lipper, Morningstar, Standard & Poor's, Fitch, etc.). These rating agencies even play an important role in the current financial regulatory framework on different continents. While these ratings are popular and even seem to determine to some extent the in- and outflow of investments in MF, these rating agencies' methodologies remain somewhat controversial, especially since the 2007 financial crisis. The ratings offered by these rating agencies have been rather intensively investigated in terms of their reliability. For instance, one rather classic study is Blake and Morey (2000) who examine the Morningstar rating in terms of its predictive power for US domestic equity MF performance: they report little evidence that Morningstar's top-rated MF outperform the second and third rated funds.

In recent years, the successful non-parametric frontier estimation methodologies from production theory have been gradually and partially transposed to the analysis of a variety of financial topics. Sengupta (1989) is to our knowledge the first to introduce an efficiency measure into a basic mean-variance (MV) portfolio model. The seminal article proposing an efficiency measure in a MF rating context is Murthi et al. (1997). But, it is probably the Morey and

[☆] We thank three referees for their most constructive comments.
[*] Corresponding author
Email addresses: olivier.brandouy@u-bordeaux4.fr (Olivier Brandouy), k.kerstens@ieseg.fr (Kristiaan Kerstens), ignace.vandewoestyne@kuleuven.be (Ignace Van de Woestyne)

Morey (1999) article proposing both a mean-return expansion and a risk contraction function that triggered a series of new developments in the use of efficiency measures in portfolio theory and in MF rating in particular.

Examples of these rather recent developments in portfolio theory include Briec et al. (2004), Lamb and Tee (2012), among others. The key advantage to the use of efficiency measures is that performance can be gauged over multiple dimensions rather than just over bi-criteria models based on mean return and some risk measure (e.g., variance). For example, Briec et al. (2007) develop mean-variance-skewness (MVS) portfolio models aiming at maximising return and skewness while minimising variance (see also Joro and Na (2006)). More recently, Branda and Kopa (2014) develop a relation between these type of models and second-order stochastic dominance (SSD) tests.

These developments have also led to a burgeoning literature on frontier-based MF rating published in a variety of outlets and covering several MF types (ethical, hedge funds, pension funds, etc.). For example, the seminal contribution of Murthi et al. (1997) employs return as an output to be magnified and risk as well as several transaction costs as inputs to be reduced: the performance of each MF is measured with respect to a piecewise linear frontier established on the universe of funds under consideration. This innovative article was quickly followed by similar models in the literature: e.g., McMullen and Strong (1998) or Premachandra et al. (1998) (see Glawischnig and Sommersguter-Reichmann (2010) for an early overview).

Finally, starting with the seminal article of Alam and Sickles (1998) there has emerged a small track in the literature providing frontier-based asset selection methods that can be integrated in portfolio models. The basic idea is to develop a link between the growth in productive performance and the evaluation of this relative frontier performance in the financial markets. Examples of these studies include, e.g., Nguyen and Swanson (2009), Edirisinghe and Zhang (2010) or Pätäri et al. (2012).

Since the advent of modern portfolio theory, a giant literature on portfolio performance gauges has emerged using total-risk foundations (e.g., the standard deviation or variance of returns). For instance, one classic is the Sharpe ratio or the reward-to-volatility index. Among the wide variety of alternative financial performance indexes, the Sortino, Treynor, Kappa and Omega ratios seem to enjoy some popularity.[1] In addition, each of the above listed MF rating agencies has come up with some rating system of its own. Therefore, our analysis employs the Sharpe ratio as well as the listed alternatives along with the Morningstar rating -a representative for the MF rating agencies- as traditional financial performance gauges.

While it is evident that a MV utility-maximising agent should target a portfolio with the highest reward-to-risk ratio (i.e., a tangency or maximum Sharpe ratio portfolio), few operational procedures in fact guarantee an investor to be positioned on the MV frontier. Since the widely used market-cap weighted indices are known to provide inefficient risk-return trade-offs, a recent research stream has aimed at making hitherto inefficient benchmarking indices efficient (for example, Clark et al. (2011)). Independent of the above frontier-based literature applied to financial topics, some authors in finance have introduced relative performance measures using some portfolio frontier as benchmark. For

[1] A recent book surveying these and more recent developments on financial and portfolio performance indicators is Bacon (2008).

instance, Cantaluppi and Hug (2000) propose an efficiency ratio relative to the MV efficient frontier very similar to Sengupta (1989). These same authors contest the rather arbitrary and non-frontier nature of most traditional financial performance gauges defining performance with respect to some other, supposedly relevant, portfolio or index.

However, it is prudent to conclude that the use of efficiency measures evaluated relative to some portfolio frontier remains a marginal idea at best in the current mainstream financial literature. This is a bit surprising in view of the important number of contributions in the behavioural finance literature documenting the biases (disposition effect, endowment effect, etc.) that prevent both non-professional and professional investors from strictly adhering to the ideal models on optimal investing (Shleifer (2000) is an early source, while Aggarwal (2014) is a recent update).[2] Just as in a consumption and production context, the use of efficiency measures within non-parametric frontier models in finance allows to document who sticks to these idealised financial models and who fails to adhere to these models and by how much. Being capable to document this heterogeneity in financial performance can be a first step to improve our understanding of the underlying causes.

The purpose of this article is then to offer the first detailed backtesting analysis of these frontier-based MF ratings compared to the Morningstar rating on the one hand (representative for the MF rating agencies), and some traditional financial performance measures on the other hand. The originality of our approach is threefold. First, we evaluate all MF ratings with respect to a traditional MV portfolio setting, such that frontier-based ratings with their capability to summarise multidimensional performance are tested with respect to the workhorse of modern portfolio management. One may speculate that frontier-based multidimensional performance ratings may have some natural advantage to improve selection for higher-order moment portfolios, but we think a first critical test for them to pass is to assess whether they are any good in the traditional MV portfolio setting.

Second, the portfolio backtesting strategy has been designed to make a minimum of assumptions. In line with the use of non-parametric frontier ratings, the portfolio allocation decisions are based on a rather recent non-parametric MV portfolio frontier model proposed by Briec et al. (2004) starting from an Equally Weighted Portfolio (EWP) composed of the 10, 20 or 30 best rated MF according to either a traditional or a frontier-based MF rating. This methodological choice guarantees making minimal assumptions to test for the usefulness of non-parametric frontier ratings in selecting promising MF compared to traditional performance measures, which is our key research question. The backtesting strategies that are in common for all performance measures used to select MF are then evaluated in terms of either terminal values (with and without transaction costs) or traditional performance measures. These are common assessment tools that guarantee a neutral environment to assess our key research question. Having described the guiding principles, we refer to the main text below for more details on the backtesting setup.

Third, the backtesting period is consciously selected to coincide with the financial crisis and therefore provides one of the harshest periods to test for the capabilities of established and new methodologies alike. To the best of our knowledge, this is probably the first extensive backtesting analysis focusing on the relative merits of backward looking

[2]This literature has also led to attempts to develop guidelines to exploit the imperfectly rational market participants (see, e.g., Montier (2007)).

traditional and frontier-based performance rating tools in predicting future MF performance within a backtesting methodology that makes minimal assumptions to create a level playing field.

The structure of this contribution is as follows. The next Section 2 offers a succinct overview of these frontier-based MF rating models. The third section describes the data employed in some detail. The details of the backtesting setup as well as the MV portfolio frontier model employed in this backtesting strategy are outlined in the next Section 4. The empirical results are presented in Section 5. Section 6 concludes the analysis and suggests a further agenda.

2. Frontier-based mutual funds rating models: Classification and Selection

Probably in view of widespread criticisms of traditional financial performance measures, several authors have introduced non-parametric frontier methods to assess MF performance. Intuitively, such non-parametric frontiers can envelop the observations of any multi-dimensional choice set and position each of these observations relative to the boundary of the choice set using some efficiency measure. In a MF context, the use of frontier or extremum estimators allows rating the performance of each MF along a multitude of dimensions instead of using just some combination of two dimensions as in most financial performance ratios (e.g., mean and variance solely). For example, the seminal Murthi et al. (1997) article specifies return as an output to be increased and risk as well as a series of transaction costs as inputs to be decreased, and the performance of each MF is measured relative to a piecewise linear frontier based on the MF universe considered. This literature on frontier MF rating has since then further developed and it has introduced a wide range of variations on this basic model. In particular, extensions have been proposed to the evaluation of pension funds, ethical MF, and hedge funds, while lower and/or upper partial moments have been utilized instead or combined with ordinary moments, among others. An up-to-date and fairly comprehensive review of this burgeoning literature is found in Glawischnig and Sommersguter-Reichmann (2010).

One potentially useful way to summarise the literature hitherto is to distinguish between several modeling approaches: (i) Models transposed from portfolio theory (see Morey and Morey (1999), for instance), (ii) Models transposed from production theory eventual in combination with some traditional financial performance measure (examples include Murthi et al. (1997) or Haslem and Scheraga (2003)), and (iii) Hedonic price models (a new proposal launched in Kerstens et al. (2011b)).

Instead of extensively discussing each of these modeling approaches, we offer some arguments to narrow down the number of potential models worthwhile considering. The new multi-moment and -period approach to portfolio analysis (case (i)) faces a fundamental difficulty to rate MF: even for small classes of MF the computational burden may be extremely high when adding several higher moments and/or time horizons. This computational problem inhibits the practical use of this approach. Focusing on the particular case of production models combined with some traditional financial performance measure(s) (case (ii)), we highlight one major problem of interpretation: when such a frontier model combines a traditional financial performance measure and some additional variables, what does the efficiency measure mean in such a setting? By way of example, Haslem and Scheraga (2003) define a frontier model

with the Sharpe index as an output combined with a series of input dimensions. While the usefulness of a frontier-based efficiency measure summarising some performance related MF information (typically based on some moments of the returns distribution, entry and exit fees, etc.) is widely acknowledged, it is hard to interpret an efficiency measure that also incorporates one or more traditional financial performance measures that are not frontier-based by conception.

Kerstens et al. (2011b) launch a new proposal to analyse MF via hedonic price models by analogy to the characteristics' approach to heterogeneous consumer goods. In effect, these authors argue that MF can best be interpreted as financial products for which the investor pays a variety of fees (entry and exit loads, among others) to have access to a managed fund whose price distribution is characterised by its moments. The number of moments that happen to matter is purely empirically determined via a nested testing approach applied to successive non-parametric hedonic price-qualities frontiers (see Kerstens et al. (2011b) for details). While this estimation of price-qualities functions and frontiers is rather common in consumer analysis and marketing, this approach is not that widespread in finance (e.g., Heffernan (1992)).

Apart from these arguments, we follow Kerstens et al. (2011b) who list a variety of specification issues that have largely been ignored in the existing frontier MF rating literature. Basically, these authors distinguish two main issues: (i) the choice of an efficiency measure, and (ii) the specification of the model linking the different dimensions involved in the MF frontier. First, these authors provide theoretical arguments for the use of the shortage function as an efficiency measure compatible with general investor preferences. In particular, this shortage function is compatible with a mixed risk aversion preference structure: odd moments need to be increased, while even moments need to be reduced. Furthermore, a slight variation on this shortage function developed in Kerstens and Van de Woestyne (2011) handles negative data values that can occur in financial applications, while maintaining a proportional interpretation (which is convenient for practitioners). Until then, this frontier MF literature employed less general efficiency measures.

Second, the same authors identify three specification issues for non-parametric frontier models to gauge MF: (i) nature of returns to scale; (ii) inclusion of higher moments and cost components; and (iii) convexity or not. These issues are analysed in the article in view of two theoretical frameworks that can guide the modeling of MF performance: portfolio theory and hedonic price theory. Summarising their analysis, these authors argue convincingly with respect to (i) that the most relevant returns to scale assumption when assessing MF with frontier models is to impose flexible (variable) returns to scale. With regard to (ii), the same authors plea to distinguish essentially between the return characteristics of a MF's share price and the shareholder transaction costs related to the buying and selling of MF shares above the net asset value per share and the expenses for MF administration and portfolio management. These authors focus on the main statistical characteristics of MF return distributions and systematically test which of the classical or robustified moments need to be included. In so doing, it is clear that one analyses MF from the viewpoint of the individual investor, not from the perspective of the mutual fund delegated manager.[3] This contrasts

[3]The limited information possessed by individual investors explains the intermediary spread fee in the market (see Brennan (1995)).

sharply to the practice of selecting some ad hoc combination of multiple variables -without any evident rule for their selection- whose frontier benchmarking yields a single aggregate efficiency score. Finally, in relation to (iii), while most non-parametric frontier articles measuring MF performance impose convexity, these authors put forward some reasons to also consider non-convexity.

In brief, based on this discussion we select both the convex and non-convex variable returns to scale (VRS) models when developing our empirical research strategy. Assuming the set of n MF under evaluation is indexed by j, ($j = 1, \ldots, n$), each MF is characterised by m input-like values x_{ij}, ($i = 1, \ldots, m$) and s output-like values y_{rj}, ($r = 1, \ldots, s$). When MF $o \in \{1, \ldots, n\}$ needs to be evaluated, then its inefficiency can be determined by the shortage function resulting from the following mathematical programming problem:

$$\begin{aligned} \max \lambda \ \text{ s.t.} & \sum_{j=1}^{n} y_{rj} z_j \geq y_{ro} + \lambda |y_{ro}|, \qquad r = 1, \ldots, s, \\ & \sum_{j=1}^{n} x_{ij} z_j \leq x_{io} - \lambda |x_{io}|, \qquad i = 1, \ldots, m, \\ & \sum_{j=1}^{n} z_j = 1, \lambda \geq 0, \\ & \forall j = 1, \ldots, n : \begin{cases} z_j \geq 0 & \text{under convexity,} \\ z_j \in \{0, 1\} & \text{under non-convexity.} \end{cases} \end{aligned} \tag{1}$$

Both models project the benchmarked MF o in the direction $g = (-|x_{1o}|, \ldots, -|x_{mo}|, |y_{1o}|, \ldots, |y_{so}|)$, whereby all output-like values y_{ro}, ($r = 1, \ldots, s$), and input-like values x_{io}, ($i = 1, \ldots, m$), are simultaneously increased and decreased in proportion to their initial values respectively. Furthermore, λ indicates the amount of inefficiency, whereby an efficient MF obtains a zero-valued shortage function (λ=0). Note that model (1) results in a linear programming (LP) problem under convexity and a mixed integer programming (MIP) problem under non-convexity.

Before putting these models to a test in the empirical section, we specify the data and the backtesting framework.

3. Sample description

For the empirical analysis in Section 5, we use the Morningstar Direct database. First, we extract a homogeneous set of 814 open-end MF, all belonging to the large caps European universe. To be more precise, 90.8% of MF belong to the Eurozone, while 9.2% are based in the UK. For these MF, we collect 156 weekly returns from 9 October 2005 to 2 October 2011. Most of these MF are euro-denominated (81.6%), about 10.4% is expressed in Pounds, and the remainder are mostly denominated in Danish, Swedish or Norwegian krone. In terms of the Morningstar classification system, the mix between Value, Blend and Growth MF is 41.6%, 46.8% and 9.6% respectively.[4] Finally, all selected MF survived during the period considered: thus, no issue of survival bias emerges in our analysis. We create a very

[4]This information is unavailable for 0.02% of the sample.

harsh testing environment on purpose by computing our ratings over a market period ranging from 2005 to 2011, but by backtesting all strategies only over the years 2008-2011, one of the worst financial crisis periods ever.

Second, to run our rating methodology, we need to associate a variety of costs generated by the MF activity and its raw returns. We have downloaded for the 814 funds the front load (entry fee), redemption load (exit fee), and the annual report net expense ratio (NER). The front load and redemption load are examples of shareholder fees and are fixed throughout the whole period under evaluation. The annual report NER reflects the actual fees charged during a particular fiscal year and proxies the total annual fund operating expenses. Clearly, this information is only available for each particular year.

Third, since some of the backtesting models (see infra) involve Morningstar ratings, we extract for the period October 2008 till September 2011 the three year Morningstar rating on a monthly basis. Fourth, since some of the backtesting models (see infra) make use of traditional financial performance measures, the monthly values for the Sharpe, Sortino, Omega, Treynor and Kappa ratios are also extracted for the same period. Fifth, backtesting also requires prices for the individual MF. Daily prices are extracted from Morningstar Direct for the selected MF from 1 January 2007 till 1 December 2011. These price data have been converted to Euros when needed at the on-going exchange rate. These price data are used to compute the first four moments of the return distribution.

Analysing the characteristics of the return distribution for the sample consisting of 814 MF over the whole period (9 October 2005 - 2 October 2011), Table 1 reports descriptive statistics on the first four moments of this population over the entire time period. The period is characterised by a low average rate of realised return of 7.10^{-4} on a weekly basis (which corresponds to nearly 3.6% on a yearly basis). The standard deviation or volatility is relatively high for these financial instruments: a variance of 9.10^{-4} on a weekly basis, which translates into a standard deviation of 21.63% on a yearly basis. The average skewness is negative (-0.86) which indicates that the return distribution is left-skewed. One can further notice a substantial average excess kurtosis equal to 6.43. This period has been particularly agitated due to the financial crisis exploding in 2008: it yielded a negative trend, a high volatility regime, negative skewness and a substantial excess kurtosis.

Concerning the fees charged to the investors, one can notice several things in the second part of Table 1. First, there is a wide and rather asymmetric distribution of front loads and redemption fees as testified by the large interquartile range and the divergence between mean and median, even though most of the redemption fees remain close to zero. Second, relatively speaking, the NER has a relatively small and less asymmetric distribution. Third, one can notice a rather substantial increase in the average NER in especially the year 2009, whereby this effect is mitigated later on.

TABLE 1 ABOUT HERE

Figure 1 and Figure 2 propose violin-plots to summarise this whole distribution for each of the four moments and fees in Table 1. Violin plots start with a box plot (with a marker for the median and a box indicating the interquartile range) and add a rotated kernel density plot to each side of this box plot. Striking features are twofold. First, the interquartile ranges of skewness and kurtosis are quite large. Second, the distributions of mean but especially variance

and kurtosis are quite asymmetric and have long tails. As to the investor fees, front loads, redemption fees and NER clearly experience a slightly asymmetric bimodal, a very asymmetric, and a mildly asymmetric distribution, respectively. The latter NER distributions becomes more asymmetric in between 2008 and 2009.

FIGURES 1 ABOUT HERE

This static picture so far hides a lot of variation. Indeed, the crisis has obviously had an important impact on the different levels of these four moments. In Figure 3 we report the values for the grand mean, grand standard deviation or volatility, grand skewness and grand excess kurtosis estimated for all 814 mutual funds in our sample over a sliding window of one year. One clearly sees that subsequent to the collapse of Lehman-Brothers (marked by a vertical red dashed line), markets entered into a quite agitated period for a while characterised by high volatility, a very negative skewness, and positive kurtosis.

FIGURE 3 AND 2 ABOUT HERE

We now move to the comparison of the computed frontier-ratings with regard to the other rating techniques or indicators collected in this study. Each of the 814 mutual funds in the sample receives 36 times 12 ratings or performance measures in our protocol: one per month over the 3-year backtesting period (yielding 36 ratings) and 12 ratings in total (both convex and non-convex MV, MVS, MVSK frontier ratings, and another six traditional financial performance measures (in casu, Morningstar, Sharpe, Sortino, Omega, Treynor and Kappa)). For each of the 814 mutual funds, we have computed a Kendall rank correlation among these 12 ratings, which delivers a hyper-cube of 814 times 12 times 12 dimensions. Then, we have aggregated the Kendall rank correlations using a simple arithmetic mean to report their overall degree of concordance in ranking. The latter is reported in Figure 4.

The key correlations results from Figure 4 can be summarised as follows. First, each family of rating (frontier vs. traditional) exhibits a strong internal consistency. All the correlation coefficients within the same family are highly positive and significant. Second, when one moves to the inter-family comparisons, the coefficient of correlation becomes negative. This is due to the fact that frontier ratings indicate an inefficiency while the traditional financial rating signal a positive performance.[5] Third and finally, the Morningstar ratings (denoted "Stars") present some particularities relative to all other ratings. Firstly, the Morningstar rating has a relatively strong negative correlation with the frontier ratings (around -0.43 to -0.44). Secondly, a consistently lower positive correlation with the other traditional financial ratings is obtained (around 0.11 to 0.12). This result suggests that the family of frontier ratings is -perhaps surprisingly- more similar to the Morningstar system than any of the other traditional financial measures.

FIGURE 4 ABOUT HERE

[5] When the shortage function becomes non-zero and increases in magnitude this means that the MF becomes less efficient. By contrast, when the Sharpe ratio, by way of example, increases then the performance of the MF improves.

4. Backtesting strategy and portfolio frontier allocations

To explore the potential benefits of the frontier-ratings presented previously, we adopt a comparative approach based on a backtesting methodology. Backtesting consists in running fictitious investment strategies using historical data so as to duplicate what could have been done by MF managers had they actually adopted these strategies in the past. Examples of such a backtesting approach are in DeMiguel et al. (2009) or Tu and Zhou (2011).

As stated in Section 1, we on purpose design a portfolio backtesting strategy making minimal assumptions. In line with the use of non-parametric frontier ratings, the portfolio allocation decisions are also based on a non-parametric MV portfolio frontier model starting from an EWP composed of the 10, 20 or 30 best rated MF according to either a traditional or a frontier-based MF rating. The backtesting strategies that are in common for all performance measures employed to select MF are evaluated in terms of either terminal values (with and without transaction costs) or traditional performance measures, which are very common and basic assessment tools. This methodological choice makes minimal assumptions to test our basic research question in a neutral environment: how useful are non-parametric frontier ratings in selecting promising MF compared to traditional performance measures?

Having summarised the basic philosophy, we now develop this backtesting strategy in great detail. Thereafter, we explain the use of the non-parametric MV portfolio frontier model.

4.1. Backtesting setup: The details

We basically investigate 12 variations of the same investment policy. Every strategy starts with the same capital of 1 monetary unit. This initial capital is invested using the specific policy defined within the strategy: it consists in selecting each rebalancing period the m 'best performing' MF in the investment universe so as to obtain an EWP.[6] From the position of this EWP in MV space (see Figure 5), an optimal portfolio is identified using the position dependent shortage function (2) as described in the next subsection. Since in each period the universe only consists of those MF selected for the EWP, the shortage function yields a projected optimal portfolio that is also composed of these same underlying MF. Thus, this additional optimisation process only modifies the weights of the MF selected in the EWP. It does never modify the set of MF considered. Thus, the strategy consists in investing in this optimal portfolio each time it is computed, thereby creating a kind of super fund. This super fund requires rebalancing the initial portfolio on a regular (i.e., monthly) basis.

Note that we have estimated the covariance matrices (as well as the co-skewness and co-kurtosis tensors) with regard to the 10, 20 or 30 best rated MF selected for the backtesting exercise. In doing so, we avoid matrix inversion problems that may eventually alter the optimisation process, and we mitigate the well known intrinsic instability as well. To be explicit, let m be the number of assets and n the number of observations, then in any of the backtesting exercises: $n \gg m$.

[6] The EWP consists in a naively diversified portfolio whereby any of the constituting assets receives an equal weight in the total amount invested irrelative to its own capitalisation. For this reason, the EWP is sometimes denoted the $1/n$ portfolio.

FIGURE 5 ABOUT HERE

Clearly, the MF selection process differentiates the 12 strategies, since the best performing MF are identified according to some performance measure. The 12 scenarios are listed in the first column of Table 2. These scenarios can be either a particular (convex or non-convex) frontier rating model (6 options), or a more traditional ranking system (6 options). First, the notation indicates which frontier rating model is used for ranking the MF to select the *m* best ones. This can be done using a convex (subscript '*c*') or a non-convex (subscript '*nc*') frontier rating model focusing on the first two (MV), three (MVS), or four moments (MVSK). For example, MVSK_c refers to the convex frontier model with expected returns, variance, skewness and kurtosis selected. Note that the loads/fees are common to all frontier models and are therefore ignored in the notation referring to a particular strategy. Second, apart from these frontier rating systems, six more traditional indicators are considered as well. In particular, we include the three year Morningstar ('Stars') rating, and the traditional Sharpe (1966), Sortino (see Sortino and Van der Meer (1991)), Omega (see Kazemi et al. (2004)), Treynor (1965), and Kappa (see Kaplan and Knowles (2004)) performance ratios.

Note that in the backtesting scenarios a selection of the 10, 20 or 30 best open-end funds is considered.[7] In the case of ties (e.g., in the Morningstar ratings or particularly when using non-convex frontier models) MF are randomly selected among the tied observations.

Using three years of data to obtain our first frontier-based rankings, we start backtesting from the 1st of October 2008 onwards. The first investment decision being made on that date, this decision is repeated each month thereafter with an updated set of ratings to select the best MF. Thus, we use a rolling window of three years with a step of a single month to compute the frontier ratings. This rating and the ensuing investment process based on projecting an EWP onto the MV frontier is repeated 36 times (months) till the end of October 2011. At this final date, we end up for each portfolio strategy with a complete historical track record of 36 monthly valuations.

The performance of all these backtesting scenarios is first and foremost gauged by evaluating and ranking the realised terminal value starting with a capital of unity, with and without transaction costs. Transaction costs are strictly proportional to the market capitalisation triggered by the portfolio re-composition (selling and buying). The proportional rate is determined by the 'front load' (entry fee) and 'redemption load' (exit fee) available in our database for each individual MF.[8] In addition, two representative traditional performance measures in finance are computed over the 36 monthly valuations as performance gauges. The Sharpe ratio is traditionally conceived as suitable for the MV world, while the Omega ratio is supposedly capable to assess a non-normal world.

[7] We restrict our analysis to this range because super funds combining 20 diversified portfolios is usually enough to obtain a satisfactory diversification in MV.

[8] We neglect other transaction costs (e.g., brokerage fees). Note also that the included transaction costs (TC) are not explicitly integrated in the portfolio optimisation process over time, but these are simply considered as fees paid in addition to the invested capital. Thus, TC are simply summarised at the end of each step in the backtesting process.

4.2. Frontier-based MV portfolio models

We now briefly describe the position dependent shortage function in a portfolio context. This shortage function is first introduced in Briec et al. (2004) with respect to a MV universe. Introducing some more notation, a portfolio consisting of n MF available in the financial universe can be considered as a vector of weights $x = (x_1, \ldots, x_n)$ indicating the individual proportions of each MF in the portfolio. By definition, $\sum_{i=1}^{n} x_i = 1$ and we ignore the possibility of shorting and operate in the absence of a risk-free rate. All MF in the financial universe are characterised by their raw returns registered over a given time window. From this information, the expected return vector and the covariance matrix can be derived. Based on the latter and the optimal portfolio weights, the expected return $\mathrm{Ret}(x)$ and variance $\mathrm{Var}(x)$ for portfolio x can be computed.

Consider a portfolio x_o under evaluation. Then, the position dependent shortage function identifies an inefficiency value λ for x_o obtained from solving the following non-linear optimisation model:

$$\begin{aligned} \max \lambda \ \ \text{s.t.}\, &\mathrm{Ret}(x) \geq \mathrm{Ret}\, x_o + \lambda |\,\mathrm{Ret}(x_o)|, \\ &\mathrm{Var}(x) \leq \mathrm{Var}\, x_o - \lambda\, \mathrm{Var}(x_o), \\ &\lambda \geq 0, \sum_{i=1}^{n} x_i = 1, \forall i = 1, \ldots, n : x_i \geq 0. \end{aligned} \tag{2}$$

The position dependent shortage function results in an inefficiency value λ. Similar with the shortage function introduced in model (1), the portfolio x_o is more efficient if its inefficiency value is closer to zero. Moreover, the optimal portfolio x corresponding with the optimal value of λ for x_o is located on the corresponding MV-frontier. Finally, it can handle negative data values for returns whenever these occur. Therefore, the use of the shortage function in portfolio analysis is consistent with its use in MF evaluation and selection (see (1)) and it is compatible with more general investor preferences (see supra).

Notice that this basic MV model provides the foundation for a lot of extensions. For instance, Briec et al. (2007) adapt this model to the MVS world, while Briec and Kerstens (2010) show that this adaptation can be generalised to an arbitrary multi-moment universe. For another example, Kerstens et al. (2011a) discuss methods to visualize the MVS-frontier, while Briec et al. (2013) explore the relation between this shortage function approach and an alternative, more popular polynomial goal programming approach due to Lai (1991), among others, in terms of the same MVS visualization. Empirical applications of this portfolio approach based on the shortage function are found in Jurczenko and Yanou (2010), Lozano and Guttiérez (2008), or Massol and Banal-Estañol (2014), among others.

5. Empirical backtesting results

Table 2 contains several performance indicators for the different backtesting portfolio strategies: two terminal values with or without transaction costs (TC), and two traditional financial ratios (Sharpe and Omega).[9] Terminal

[9]Note that while the portfolio strategies have been phrased in terms of Sharpe, Sortino, Omega, Treynor and Kappa ratios, by lack of space we limit the evaluation of all 12 backtesting scenarios to just a selection of two representative ratios.

wealth is expressed as a percentage of unit initial wealth. The size of the different super funds is reported between brackets in the second line in the table header. The different names of the backtesting portfolio strategies can be found in the first column (see supra). Note that the first six rows relate to frontier ratings, while the last six rows relate to more traditional performance measures. Below each of the four performance indicators, one finds a rank in descending order computed over all possible strategies within that column. The two last columns of the table report the harmonic mean (HM) and harmonic standard deviations (HSD) of the ranks obtained by each strategy in the same row.[10]

TABLE 2 ABOUT HERE

The first key observation is that the frontier-based strategies largely outperform strategies based on more traditional indicators. This is first of all evident for the terminal value without TC: frontier-ratings always guarantee a sure gain, while the others only do so in 8 out of 18 cases (whereby 'Stars' also systematically grant gains). In addition, for each super fund size and for each performance indicator they occupy the first six positions in 60 out of 72 cases. For the 12 remaining cases, the 'Stars'-driven strategies end up in the top half in 8 cases, while Sharpe- and Omega-driven strategies manage to do this only twice each.[11]

Second, the frontier-based strategies imply most of the times lower overall transaction costs compared to the strategies based on traditional indicators. Transaction costs for each strategy and for each sample size can be computed as the difference in terminal value without and with TC. Using these values, the grand mean and standard deviation of transaction costs for the frontier-based strategies amounts to 20.78% and 3.24%. The same statistics for the strategies driven by the traditional performance indicators yield a grand mean and standard deviation of 31.04% and 11.57%. This indisputable result at the aggregate level could be seen as a first indication of a greater stability in the intersection of succeeding subsets of best MF over time for frontier ratings.

A third observation relates to the relative coherency and consistency of all these financial ratings. While the frontier-ratings are even judged favorably by the two traditional financial ratios, the Sharpe- and Omega-based strategies only manage to appear twice each in the top half according to their own performance rating (but never in terms of the other performance measures). Thus, these traditional measures appear to be little coherent. This remark extends to the strategies based on the Sortino or the Kappa ratios which both perform poorly when gauged with a higher-moment traditional indicator (i.e., the Omega ratio). Similarly, a strategy based on a selection of the best MF using the Treynor ratio poorly performs when gauged with the Sharpe ratio. Furthermore, it is also obvious that the frontier-ratings are more consistent in their rankings across all experimental treatments under the four performance indicators. In partic-

[10] Ranks obtained by each strategy are aggregated per row using the harmonic mean (HM) or the harmonic standard deviation (HSD) of the ranks X_i: $HM = \left(n^{-1} \sum_{i=1}^{n} X_i^{-1}\right)^{-1}$ and $HSD = \sqrt{\sigma^2(X_i^{-1})} \Big/ \left(n\left(\overline{X_i^{-1}}\right)^4\right)$. For both these statistics, the lower the value the more desirable the strategy.

[11] As for the negative value of the Sharpe ratios documented in this study, note that we backtest all strategies over one of the worst financial crisis periods ever (2008-2011).

ular, these frontier-ratings all obtain harmonic mean and standard-deviation ranks below 6 (with just two exceptions: MV_{NC} for HM, and MV_C for HSD). This indicates that their rankings are both better and more stable.

Instead of judging strategies based solely on their ranking, we can also evaluate these using the entire distribution. Figure 6 offers a graphical overview of the performance per strategy by stacking a box plot for each. In each of these three Figures 6(a)-6(c) depending on whether 10, 20 or 30 MF enter into the EWP, the strategies are sorted in descending order from top to bottom with respect to the level of the mean return obtained during the backtesting simulations. The small red triangle for each strategy reports the location of this mean return and all box plots are aligned on the number 0. Thus, moving from bottom to the top implies that the mean return increases monotonously.

Three main conclusions emerge. First, the frontier based strategies as well as the 'Stars' perform better than traditional financial performance gauges. Second, the 'Stars' measure is dominated by at least three frontier-based strategies. Third, the multi-moment ratings do better than the traditional MV ratings in the convex case, though for the non-convex models this is not the case in general. Traditional convex frontier models seem somewhat to outperform the newer non-convex models on average, even if the latter have sometimes a more compact distribution.

FIGURE 6 ABOUT HERE

Finally, we have also run a hierarchical cluster analysis on the scores obtained by each backtesting strategy according to each of the four performance gauges (i.e., terminal value, terminal value with transaction costs, Sharpe and Omega ratios) as displayed in Table 2. The underlying idea is to group the strategies employed to manage these backtested portfolios over time within a series of similar clusters with respect to their descriptors (*in casu*, the set of 12 gauges). The clustering algorithm selected is the Ward (1963) minimum variance method which seeks at finding clusters that are as compact as possible.[12] Figure 7 presents the dendrogram of this hierarchical cluster analysis. Figure 8 shows the evolution of the sum of squared errors with respect to an increasing number of K-means clusters. Even though it is always somewhat arbitrary to cut the dendrogram so to obtain an optimal number of clusters, three groups of strategies tend to emerge from the analysis of Figure 8. One usually chooses the number of clusters where an elbow appears in Figure 8: in our case, the choice of 3 clusters appears evident.

These three clusters can be characterised as follows. A first cluster contains all traditional financial performance gauges, apart from the Morningstar rating. A second group is composed of the 'Stars' rating, as well as two non-convex frontier models (i.e., MVS_{nc} and $MVSK_{nc}$). The last group gathers all convex frontier based models (i.e., MV_c, MVS_c and $MVSK_c$), as well as the basic MV non-convex frontier model (MV_{nc}).

FIGURES 7 and 8 ABOUT HERE

The overall conclusions of this threefold analysis of the backtesting results are clear. First, frontier-based ratings of MF seem to offer better tools than long standing financial ratings. Second, some of these frontier-based ratings of

[12] Practically speaking, we rely on the implementation of this method as described in Murtagh and Legendre (2013).

MF even perform better than the Morningstar rating. Third, the cluster analysis even seems to indicate a structural similarity between the non-convex frontier ratings and the Morningstar rating, even though the convex frontier ratings seem to do better than their non-convex counterparts on average.

6. Conclusions

This contribution provides the first backtesting exercise comparing the recent convex and non-convex frontier MF rating models against traditional financial performance measures (i.e., Sharpe, Sortino, Omega, Treynor, and Kappa) and the three year Morningstar rating. The punchline of this analysis is rather simple: frontier-based MF ratings allow to design investment policies generating better performances than most of its competitors, and these frontier-ratings are comparatively more coherent and consistent than most traditional counterparts.

The setup we have chosen to run this exercise is as follows: among a set of 814 MF, we compare strategies based on the selection of 10, 20 or 30 best ranked MF that are used to create an EWP of MF. We employ 12 different rating systems to do so: 6 adhere to a frontier-based model, while 6 belong to the traditional performance approaches (including Morningstar). Then, we project this EWP towards the MF MV frontier using the shortage function. The resulting optimal super fund is then chosen as the target portfolio tracked over the backtesting exercise. This process is updated each month from October 2008 till October 2011 with a rolling time window of three years.

In this framework, we clearly establish an overall dominance of the strategies based on frontier-based ratings over those that exploit more classical ratings. This dominance is rather clearly established with regard to the terminal wealth (with or without transaction costs) of a fictitious investor choosing one among these 12 strategies. Frontier-based strategies tend to identify a subset of MF which seems more stable over time than the ones identified by the other strategies without a sacrifice in the level of performance.

The same conclusions can be drawn from the analysis of the Sharpe or the Omega ratios obtained by these strategies. However, in three cases (out of 72) the Morningstar rating system performs even better (obtaining the first rank among 12), but only when the subset of MF composing the super fund is very large (30 MF). However, the performance spread between Morningstar and the second rated strategy (i.e., $MVSK_c$) is almost negligible. Apart from the Morningstar rating, none of the studied strategies belonging to the traditional performance ratings obtain ranks below 6 out of 12.

This result based on ratings is confirmed in an analysis based on comparing entire distributions via stacked and aligned box plots. Furthermore, a hierarchical cluster analysis confirms the separation between the traditional financial ratings on the one hand and the frontier-based MF ratings as well as the Morningstar rating on the other hand. One may therefore prudently conclude that the Morningstar star rating may have finally found a serious contender.

Even if further extensive backtesting studies are called for to better explore the potential benefits of frontier-based ratings for MF selection, these first results are very promising and open up an exciting perspective for applications of frontier-based MF ratings for both finance researchers and investors alike. In particular, further intensive backtesting

could prove beneficial in at least the following areas: the impact of the exact starting period of the backtesting period on its results for this given sample, the effect of selecting an even larger validation period and/or an even larger holdout period, the effect of choosing more robust statistics to describe the return distribution (see Martin et al. (2010) in general and Kerstens et al. (2011b) or Yanou (2013) on L-moments and trimmed L-moments in a financial context) rather than ordinary moments, etc.

References

Aggarwal, R., 2014. Animal spirits in financial economics: A review of deviations from economic rationality. International Review of Financial Analysis 32, 179–187.

Alam, I., Sickles, R., 1998. The relationship between stock market returns and technical efficiency innovations: Evidence from the us airline industry. Journal of Productivity Analysis 9 (1), 35–51.

Bacon, C., 2008. Practical portfolio performance measurement and attribution, 2nd Edition. Wiley.

Blake, C., Morey, M., 2000. Morningstar ratings and mutual fund performance. Journal of Financial and Quantitative Analysis 35 (3), 451–483.

Branda, M., Kopa, M., 2014. On relations between dea-risk models and stochastic dominance efficiency tests. Central European Journal of Operations Research 22 (1), 13–35.

Brennan, M., 1995. The individual investor. Journal of Financial Research 18 (1), 59–74.

Briec, W., Kerstens, K., 2010. Portfolio selection in multidimensional general and partial moment space. Journal of Economic Dynamics and Control 34 (4), 636–656.

Briec, W., Kerstens, K., Jokung, K., 2007. Mean-variance-skewness portfolio performance gauging: A general shortage function and dual approach. Management Science 53 (1), 135–149.

Briec, W., Kerstens, K., Lesourd, J., 2004. Single-period Markowitz portfolio selection, performance gauging, and duality: A variation on the Luenberger shortage function. Journal of Optimization Theory and Applications 120 (1), 1–27.

Briec, W., Kerstens, K., Van de Woestyne, I., 2013. Portfolio selection with skewness: A comparison of methods and a generalized one fund result. European Journal of Operational Research 230 (2), 412–421.

Cantaluppi, L., Hug, R., 2000. Efficiency ratio: A new methodology for performance measurement. Journal of Investing 9 (2), 1–7.

Clark, E., Jokung, O., Kassimatis, K., 2011. Making inefficient market indices efficient. European Journal of Operational Research 209 (1), 83–89.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? Review of Financial Studies 22, 1915–1953.

Edirisinghe, N., Zhang, X., 2010. Input/output selection in DEA under expert information, with application to financial markets. European Journal of Operational Research 207 (3), 1669–1678.

Glawischnig, M., Sommersguter-Reichmann, M., 2010. Assessing the performance of alternative investments using non-parametric efficiency measurement approaches: Is it convincing? Journal of Banking & Finance 34 (2), 295–303.

Haslem, J., Scheraga, C., 2003. Data Envelopment Analysis of Morningstar's large-cap mutual funds. Journal of Investing 12 (4), 41–48.

Heffernan, S., 1992. A computation of interest equivalences for non-price features of bank products. Journal of Money, Credit and Banking 24 (2), 162–172.

Joro, T., Na, P., 2006. Portfolio performance evaluation in mean-variance-skewness framework. European Journal of Operational Research 175 (1), 446–461.

Jurczenko, E., Yanou, G., 2010. Fund of hedge funds portfolio selection: A robust non-parametric multi-moment approach. In: Watanabe, Y. (Ed.), The Recent Trend of Hedge Fund Strategies. Nova Science, New York, pp. 21–56.

Kaplan, P., Knowles, J., 2004. Kappa: A generalized downside risk-adjusted performance measure. Journal of Performance Measurement 8 (3), 42–54.

Kazemi, H., Schneeweis, T., Gupta, B., 2004. Omega as a performance measure. Journal of Performance Measurement 8 (3), 16–25.

Kerstens, K., Mounir, A., Van de Woestyne, I., 2011a. Geometric representation of the mean-variance-skewness portfolio frontier based upon the shortage function. European Journal of Operational Research 210 (1), 81–94.

Kerstens, K., Mounir, A., Van de Woestyne, I., 2011b. Non-parametric frontier estimates of mutual fund performance using C- and L-moments: Some specification tests. Journal of Banking & Finance 35 (5), 1190–1201.

Kerstens, K., Van de Woestyne, I., 2011. Negative data in DEA: A simple proportional distance function approach. Journal of the Operational Research Society 62 (7), 1413–1419.

Lai, T. Y., 1991. Portfolio selection with skewness: A multiple objective approach. Review of Quantitative Finance and Accounting 1 (3), 293–305.

Lamb, J., Tee, K.-H., 2012. Data envelopment analysis models of investment funds. European Journal of Operational Research 216 (3), 687–696.

Lozano, S., Guttiérez, E., 2008. TSD-consistent performance assessment of mutual funds. Journal of the Operational Research Society 59 (10), 1352–1362.

Martin, R., Clark, A., Green, C., 2010. Robust portfolio construction. In: Guerard, J. (Ed.), Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques. Springer, Berlin, pp. 337–380.

Massol, O., Banal-Estañol, A., 2014. Export diversification through resource-based industrialization: The case of natural gas. European Journal of Operational Research 237 (3), 1067–1082.

McMullen, P., Strong, R., 1998. Selection of mutual fund using Data Envelopment Analysis. Journal of Business and Economic Studies 4 (1), 1–14.

Montier, J., 2007. Behavioural Investing: A Practitioner's Guide to Applying Behavioural Finance. Wiley, New York.

Morey, M., Morey, R., 1999. Mutual fund performance appraisals: A multi-horizon perspective with endogenous benchmarking. Omega 27 (2), 241–258.

Murtagh, F., Legendre, P., 2013. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? Journal of Classification forthcoming.

Murthi, B., Choi, Y., Desai, P., 1997. Efficiency of mutual funds and portfolio performance measurement: A non-parametric approach. European Journal of Operational Research 98 (2), 408–418.

Nguyen, G., Swanson, P., 2009. Firm characteristics, relative efficiency, and equity returns. Journal of Financial and Quantitative Analysis 44, 213–236.

Pätäri, E., Leivo, T., Honkapuro, S., 2012. Enhancement of equity portfolio performance using data envelopment analysis. European Journal of Operational Research 220 (3), 786–797.

Premachandra, I., Powell, J., Shi, J., 1998. Measuring the relative efficiency of fund management strategies in New Zealand using a spreadsheet-based stochastic Data Envelopment Analysis model. Omega 26 (2), 319–331.

Sengupta, J., 1989. Nonparametric tests of efficiency of portfolio investment. Journal of Economics 50 (1), 1–15.

Sharpe, W., 1966. Mutual fund performance. Journal of Business 39 (1), 119–138.

Shleifer, A., 2000. Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press, Oxford.

Sortino, F., Van der Meer, R., 1991. Downsize risk. Journal of Portfolio Management 18 (2), 27–32.

Treynor, J., 1965. How to rate management of investment funds. Harvard Business Review 43 (1), 63–75.

Tu, J., Zhou, G., 2011. Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. Journal of Financial Economics 99, 204–215.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58 (301), 236–244.

Yanou, G., 2013. Extension of the random matrix theory to the l-moments for robust portfolio selection. Quantitative Finance 13 (10), 1653–1673.

	Returns				Fees (in percentage)				
	Mean	Variance	Skewness	Kurtosis	Front[(1)]	Redemption[(2)]	NER 2008	NER 2009	NER 2010
Min.	-0.0011	0.0002	-2.1400	1.2400	0.000	0.000	0.000	0.120	0.130
Q1	0.0004	0.0008	-1.0700	4.8700	0.200	0.000	1.180	1.210	1.202
Median	0.0006	0.0009	-0.8900	6.1600	3.625	0.000	1.595	1.630	1.630
Mean	0.0007	0.0009	-0.8640	6.4300	2.977	0.278	1.590	1.620	1.611
Q3	0.0009	0.0009	-0.6100	7.7900	5.000	0.000	1.867	1.897	1.890
Max.	0.0039	0.0026	0.5300	23.5000	10.000	6.000	7.370	5.510	5.250

(1) Front load, (2) Redemption fee.

Table 1: Descriptive statistics for all 814 MF

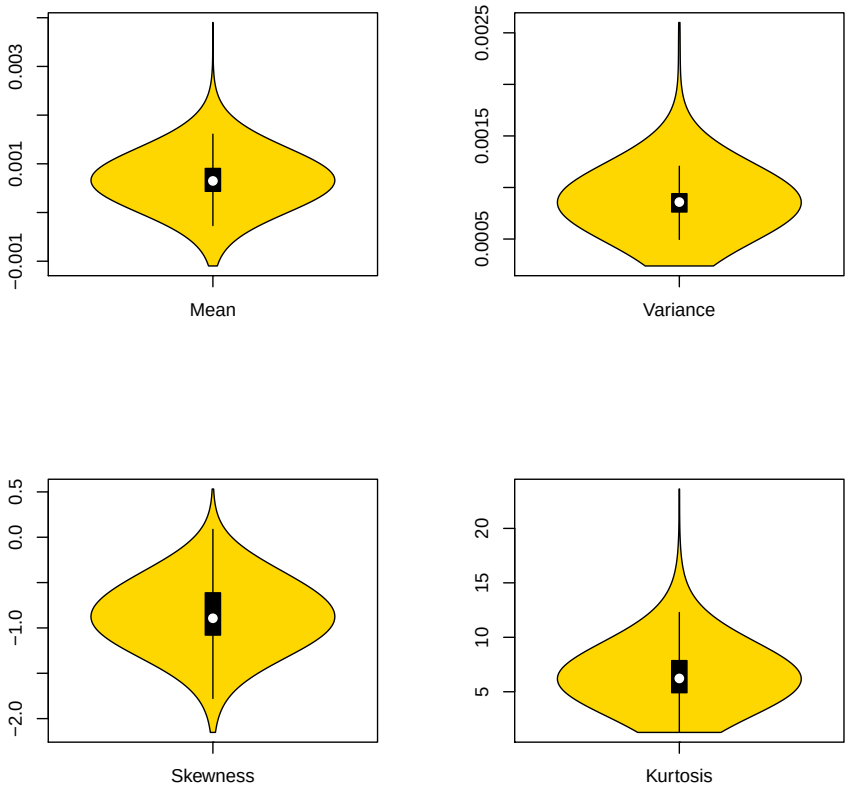


Figure 1: Violin plots of moments distribution

Figure 2: Violin plots of fees distribution

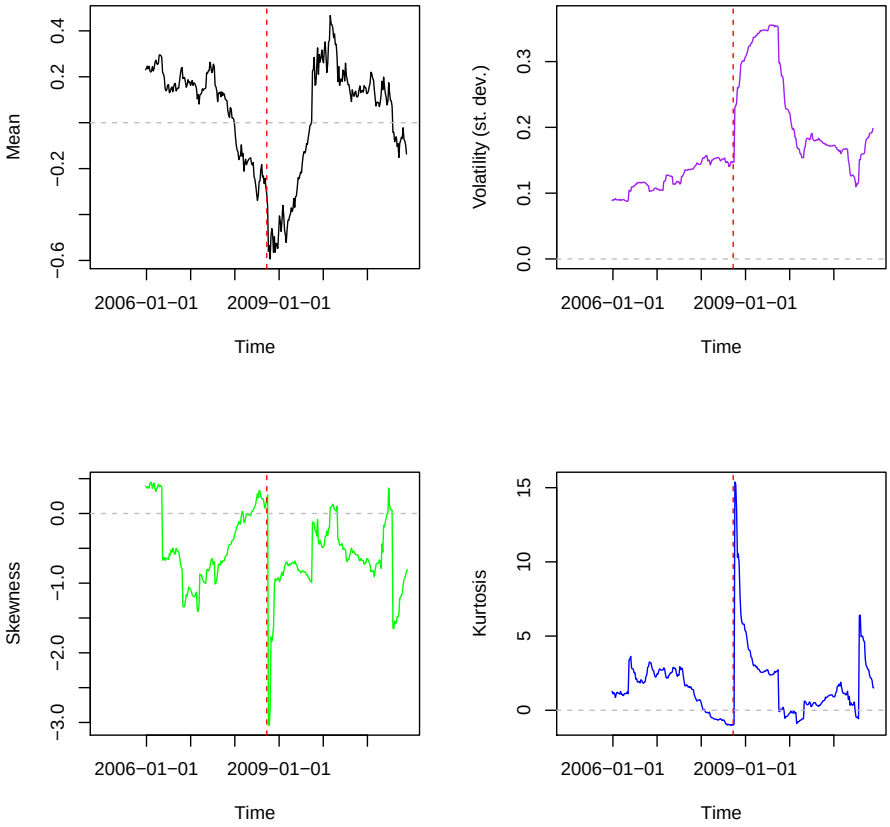


Figure 3: Evolution of moment distribution of aggregate returns over time

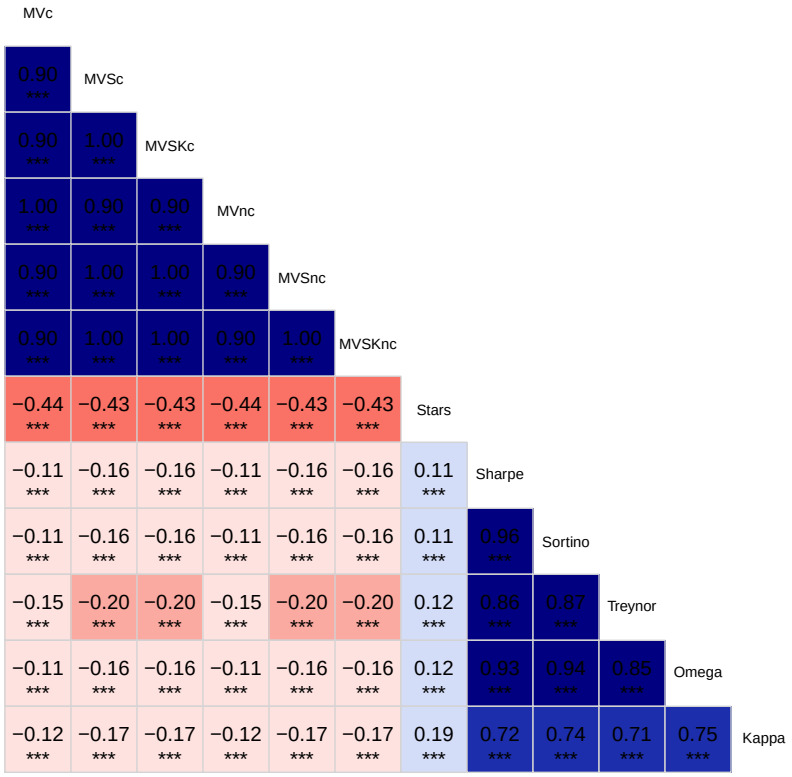


Figure 4: Correlation matrix of frontier and traditional MF ratings[†]

[†] *, **, and *** denote significance at a 5%, 1% and 0.1% level resp.

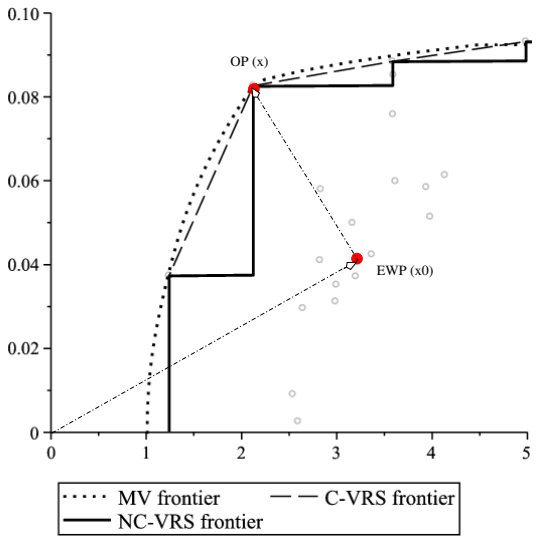
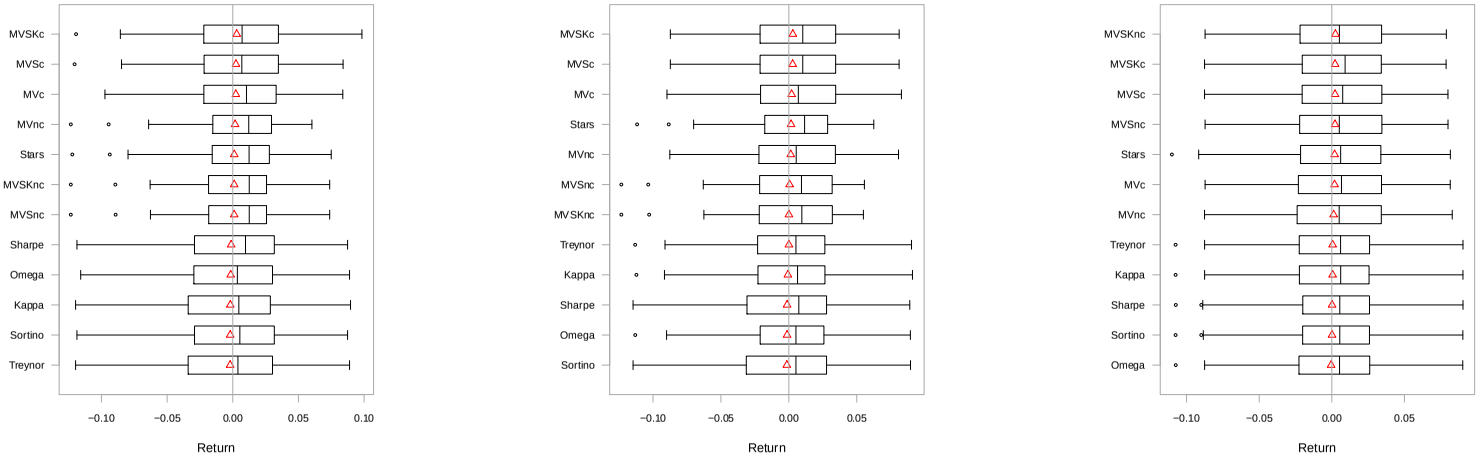


Figure 5: Projection of the EWP towards the MV frontier

	Terminal Value (no TC)			Terminal Value (with TC)			Sharpe Ratio			Omega Ratio			Rank Aggreg.	
	MF(10)	MF(20)	MF(30)	MF(10)	MF(20)	MF(30)	MF(10)	MF(20)	MF(30)	MF(10)	MF(20)	MF(30)	HM	HSD
MV_c	1.0844	1.0752	1.0680	0.8563	0.8079	0.8152	-0.0945	-0.1356	-0.1251	0.7891	0.7074	0.7264		
	3	*3*	*6*	*5*	*4*	*6*	*5*	*4*	*6*	*4*	*4*	*6*	6	7
MVS_c	1.0881	1.1038	1.0810	0.9557	0.9538	0.9022	-0.0292	-0.0322	-0.0704	0.9287	0.9221	0.8372		
	2	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	2	2
$MVSK_c$	1.1101	1.1085	1.0815	0.9297	0.9563	0.9110	-0.0430	-0.0305	-0.0643	0.8926	0.9258	0.8517		
	1	*1*	*2*	*2*	*1*	*2*	*2*	*1*	*2*	*2*	*1*	*2*	1	1
MV_{nc}	1.0639	1.0505	1.0447	0.6961	0.8843	0.8077	-0.2069	-0.0796	-0.1314	0.5629	0.8177	0.7171		
	4	*5*	*7*	*7*	*3*	*7*	*11*	*3*	*7*	*11*	*3*	*7*	7	6
MVS_{nc}	1.0383	1.0207	1.0798	0.8738	0.6788	0.8854	-0.0859	-0.2039	-0.0736	0.7929	0.5574	0.8229		
	7	*6*	*4*	*3*	*5*	*4*	*3*	*5*	*4*	*3*	*5*	*4*	5	5
$MVSK_{nc}$	1.0389	1.0029	1.0899	0.8649	0.5999	0.8822	-0.0914	-0.2608	-0.0766	0.7807	0.4764	0.8157		
	6	*7*	*1*	*4*	*7*	*5*	*4*	*9*	*5*	*5*	*10*	*5*	4	4
Sharpe	0.9576	0.9561	1.0071	0.6551	0.5806	0.6055	-0.1836	-0.2246	-0.2278	0.6188	0.5336	0.5212		
	8	*10*	*10*	*9*	*10*	*10*	*7*	*6*	*8*	*7*	*6*	*8*	8	8
Treynor	0.9319	1.0017	1.0194	0.5937	0.5842	0.6086	-0.2182	-0.2406	-0.2475	0.5540	0.5141	0.5015		
	12	*8*	*8*	*12*	*9*	*9*	*12*	*8*	*10*	*12*	*8*	*10*	10	11
Sortino	0.9340	0.9500	1.0054	0.6095	0.5873	0.5728	-0.2048	-0.2326	-0.2524	0.5799	0.5205	0.4871		
	11	*12*	*11*	*11*	*8*	*11*	*10*	*7*	*11*	*9*	*7*	*11*	12	10
Stars	1.0397	1.0595	1.0700	0.7093	0.6222	0.9145	-0.1943	-0.2708	-0.0549	0.5751	0.4313	0.8643		
	5	*4*	*5*	*6*	*6*	*1*	*8*	*11*	*1*	*10*	*12*	*1*	3	3
Omega	0.9460	0.9558	0.9834	0.6646	0.4865	0.5182	-0.1824	-0.2837	-0.3030	0.6220	0.4351	0.4126		
	9	*11*	*12*	*8*	*12*	*12*	*6*	*12*	*12*	*6*	*11*	*12*	11	9
Kappa	0.9363	0.9755	1.0194	0.6336	0.5702	0.6244	-0.1986	-0.2629	-0.2464	0.5871	0.4835	0.5073		
	10	*9*	*8*	*10*	*11*	*8*	*9*	*10*	*9*	*8*	*9*	*9*	9	12

Table 2: Performance results for 12 backtesting scenarios



(a) Universe MF(10)

(b) Universe MF(20)

(c) Universe MF(30)

Figure 6: Return distributions (stacked and aligned boxplots) for the 12 studied strategies

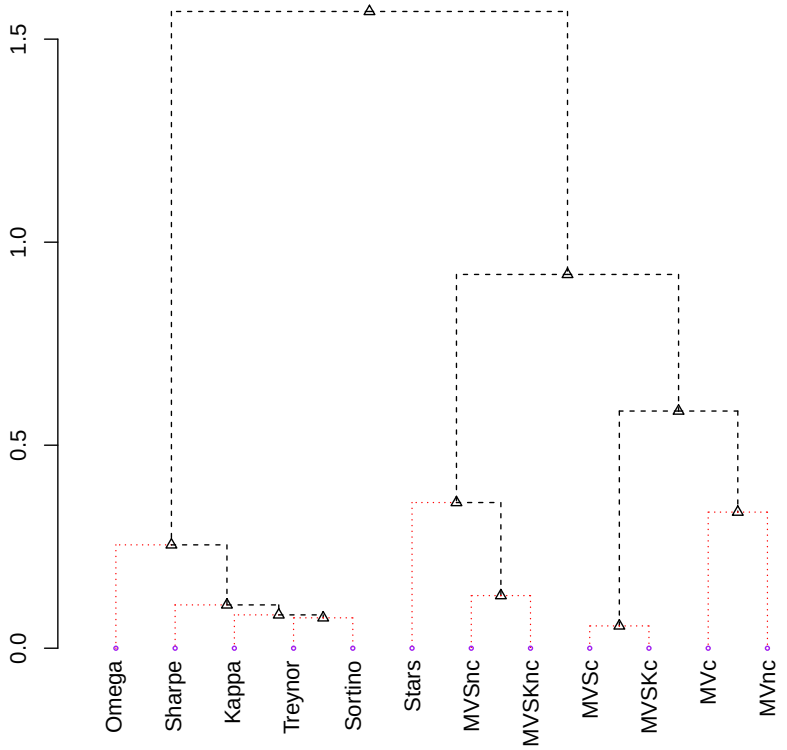


Figure 7: Dendrogram of the 12 backtesting strategies

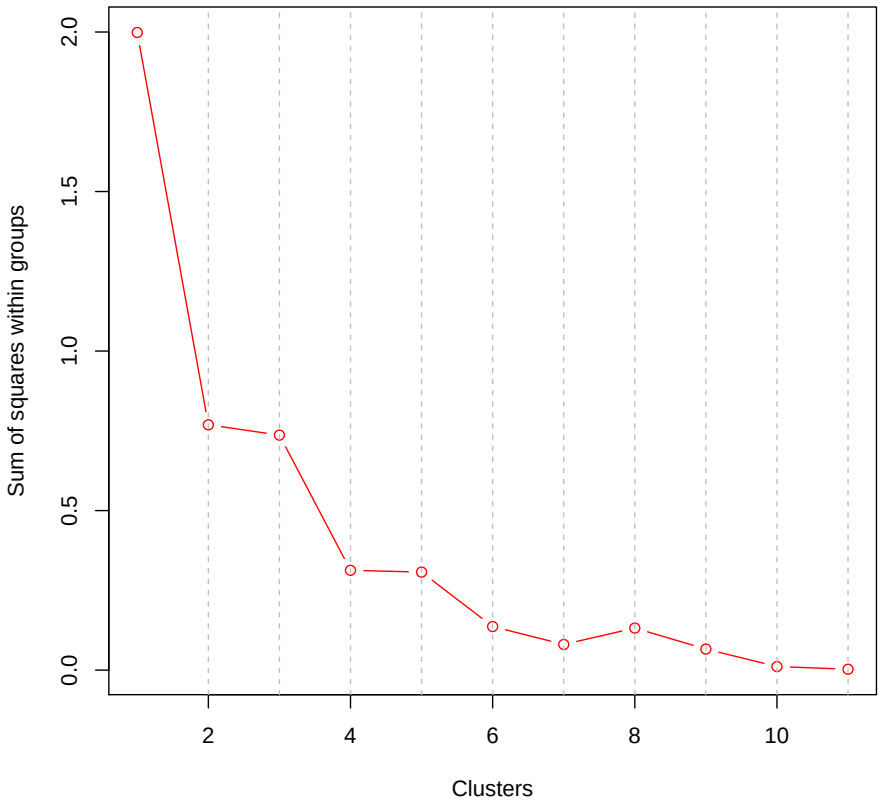


Figure 8: SSE of the K-means clustering algorithms